View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by LSE Research Online

LSE Research Online

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

J. McKenzie Alexander

Cheap talk, reinforcement learning and the emergence of cooperation

Article (Accepted version) (Refereed)

Original citation:
Alexander, J. McKenzie (2014) *Cheap talk, reinforcement learning and the emergence of cooperation*. Philosophy of Science. ISSN 0031-8248 (In Press)

© 2014 University of Chicago Press

This version available at: http://eprints.lse.ac.uk/57315/
Available in LSE Research Online: August 2014

LSE has developed LSE Research Online so that users may access research output of the School. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LSE Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain. You may freely distribute the URL (http://eprints.lse.ac.uk) of the LSE Research Online website.

This document is the author's final accepted version of the journal article. There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

http://eprints.lse.ac.uk

Cheap Talk, Reinforcement Learning and the Emergence of Cooperation

J. McKenzie Alexander
Department of Philosophy, Logic and Scientific Method
London School of Economics and Political Science

28th June 2014

Cheap talk has often been thought incapable of supporting the emergence of cooperation because costless signals, easily faked, are unlikely to be reliable (Zahavi and Zahavi, 1997). I show how, in a social network model of cheap talk with reinforcement learning, cheap talk does enable the emergence of cooperation, provided that individuals also temporally discount the past. This establishes one mechanism that suffices for moving a population of initially uncooperative individuals to a state of mutually beneficial cooperation even in the absence of formal institutions.

1. The many roads to cooperation. Explaining how cooperative behaviour — or pro-social behaviour, more generally — might emerge has become a cottage industry. Mechanisms which have been shown to work, in certain contexts, include the following: reliable signals, or the "secret handshake" (Robson, 1990); costly signals, a.k.a. "the handicap principle" (Zahavi, 1975; Zahavi and Zahavi, 1997); punishment (Boyd and Richerson, 1992; Gintis, 2000); compliance with social norms (Axelrod, 1986; Bowles and Gintis, 2004; Bicchieri, 2005); correlated interactions induced by social structure (Nowak and May, 1993; Ellison, 1993; Skyrms, 2003; Alexander, 2007); reciprocal altruism (Trivers, 1971); and group selection (Sober and Wilson, 1998). This list is by no means exhaustive.

Some of these mechanisms support the emergence of cooperative behaviour simply because the mechanism is considerably flexible it terms of what it may yield; recall, after all, that the title of Boyd and Richerson's paper is "punishment allows the evolution of cooperation (or anything else) in sizable groups". Other mechanisms have more limited scope. Social norms require people to not only

know the behaviour of others, and the underlying rule which governs their behaviour, but also other people's *expectations*. And local interaction models, although quite effective in supporting cooperation (Nowak and May, 1993), fairness (Alexander and Skyrms, 1999), and trust (Skyrms, 2001, 2003), face greater difficulty explaining the behaviour of agents in the ultimatum game.

Finally, the mechanism of costly signals seems to have the most limited scope of all the methods listed above. Why? Zahavi argued signals must be costly in order to ensure that they are reliable, or honest, for otherwise such signals could be easily forged by opportunistic agents. Yet forging signals would only be problematic in cases involving altruistic behaviour, such as the Prisoner's Dilemma or the Sir Philip Sidney game (Maynard Smith, 1991) — games where "cooperating" leaves the agent vulnerable to exploitation. In a pure coordination game, like the Driving Game, or an impure coordination game, like Battle of the Sexes, or even a trust game like the Stag Hunt, an honest signal need not be costly. In these games, receipt of an honest signal may *increase* the chance of arriving at a Pareto-optimal Nash equilibrium.

Furthermore, some have challenged whether signals need be costly in order to be effective even in cases of *altruistic* behavior. In series of three articles, Carl Bergstrom and Michael Lachmann consider the effects of costly versus costless signals in the Sir Philip Sidney game. They find that, in some cases, costly signalling can be *so* costly that individuals are worse off than not being able to signal at all; they also show that honest cost-free signals are possible under a wide range of conditions. Similarly, Huttegger and Zollman (2010) show — again for the Sir Philip Sidney game — that the costly signalling equilibrium turns out to be less important for understanding the overall evolutionary dynamics than previously considered.

In what follows, I contribute to the critique of the importance of costly signals for the emergence of cooperation, but using a rather different approach than what has previously been considered. In section 2, I present a general model of reinforcement learning in network games, which builds upon the work of Skyrms (2010) and Alexander (2007). Section 3 introduces the possibility of costless *cheap talk* into this model, as well as the possibility of conditionally responding to received signals. I then show that — in accordance with the Handicap Principle — cooperative behaviour does not emerge. However, in section 4 I show that when cheap talk and reinforcement learning is combined with *discounting the past*, that costless signalling enables individuals to learn to cooperate despite originally settling upon a "norm" of defecting.

2. Reinforcement learning in network games. If one were to identify one general trend in philosophical studies of evolutionary game theory over the past twenty years, it would be a movement towards ever-more limited models of

boundedly rational agents. Contrast the following: in *The Dynamics of Rational Deliberation*, Skyrms modelled interactions between two individuals who updated their beliefs using either Bayesian or Brown-von Neumann-Nash dynamics, both fairly cognitively demanding. In his most recent book *Signals: Evolution, Learning & Information*, Skyrms almost exclusively employs reinforcement learning in his simulations.

This strategy of attempting to *do more with less* has some advantages. For one, in real life we rarely know the actual payoff structure of the games we play. Without knowing the payoff matrix, we cannot even begin to calculate the expected best-response (to say nothing of the difficulty of trying to attribute degrees of belief to our opponents). Reinforcement learning doesn't require that agents know the payoff matrix.[1] Second, even a relatively simple learning rule like *imitate-the-best* requires knowledge of two things: the strategy used by our opponents, and the payoffs they received. Even if we set aside worries about interpersonal comparison of utilities, there is still the problem that the behavioural outcome of different strategies can be observationally the *same* — so which strategy does an agent adopt using *imitate-the-best*?[2]

Another advantage of reinforcement learning is that several different varieties have been studied empirically, with considerable efforts made to develop descriptively accurate models of human learning. Two important variants are due to Bush and Mosteller (1951, 1955) and Roth and Erev (1995). Let us consider each of these in turn.

Suppose that there are N actions available, and $p_i(t)$ denotes the probability assigned to action i at time t. Bush-Mosteller reinforcement learning makes incremental adjustments to the probability distribution over actions so as to move the probability of the reinforced action towards one. The *speed* with which this occurs is controlled by a learning parameter a.[3] If the kth action is reinforced, the new probability $p_k(t+1)$ is just $p_k(t) + a(1 - p_k(t))$. All other probabilities are decremented by the amount $\frac{1}{N-1}a(1 - p_k(t))$ in order to ensure that the probabilities sum to 1. One point to note is that this means Bush-Mosteller reinforcement learning does not take into account past experience: if you assign probability $\frac{3}{4}$ to an action, and reinforce, your probability distribution shifts

[1] Although, given enough experience and memory, an agent would be able to reconstruct at least her side of the payoff matrix.

[2] Consider the ultimatum game where an agent in the role of Receiver can take one of four actions: Accept always, Accept if fair, Reject if fair, and Reject always. (In the absence of acceptance thresholds, these are the four logical possibilities.) Suppose I make you an unfair offer and you reject. Suppose that I now notice that you did the best of all my opponents. What acceptance strategy should I use?

[3] In their original paper, Bush and Mosteller also included a parameter representing factors which decreased the probability of actions. For simplicity, I omit this.

by the same amount it would if you assigned probability $\frac{3}{4}$ to an action after a thousand trials.

Roth and Erev reinforcement learning, in contrast, is a form of reinforcement learning which take past experience into account. It can be thought of as a Pólya urn: each agent has an urn filled with an initial assortment of coloured balls representing the actions available. An agent draws a ball, performs the corresponding act, and then reinforces by adding a number of similarly coloured balls based on the reward of the act. More generally, an agent may assign arbitrary positive-valued weights to each act, choosing an act with probability proportional to its weight. This latter representation drops the requirement that the weights assigned to acts be integral-valued, which allows one to incorporate additional aspects of human psychology into the model, such as discounting the past.

To see how Roth-Erev takes experience into account, suppose that the reward associated with an act is always one, and that there are exactly two available acts. If the agent initially starts with an urn containing a red ball representing act 1 and a green ball representing act 2, then the initial probability of each act is $\frac{1}{2}$. Reinforcing act 1 will cause the urn to have two red balls and one green ball, so the new probability of act 1 is $\frac{2}{3}$. But now suppose that after 20 trials the urn contains exactly 10 red balls and 10 green balls. Reinforcing act 1, at this point, causes the probability of act 1 to increase to only $\frac{11}{20}$.

Roth-Erev reinforcement learning has some nice theoretical properties, aside from the limited epistemic requirements it imposes on agents. Consider the idealised problem of choosing a restaurant in a town where you don't speak the language. The challenge you face is the trade-off between *exploration* and *exploitation*. You don't want to settle for always eating at the first restaurant that serves you a decent meal. However, you also don't want to keep sampling indefinitely, so that you never fixate upon a single restaurant.[4] How should you learn from your experience so as to avoid both of these two errors? If you approach the restaurant problem as a Roth-Erev reinforcement learner, with the urn initially containing one ball for each restaurant,[5] then in the limit you will converge to eating at the best restaurant in town, always.[6] Because of these nice theoretical properties, I shall concentrate exclusively on Roth-Erev reinforcement learning in what follows.

Now consider the following basic model: let $P = \{a_1, \ldots, a_n\}$ be a population of boundedly rational agents situated within a social network (P, E), where E is a set of undirected edges. This network represents the structure of the population,

[4] Let us assume that the restaurant all serve a sufficiently generic cuisine so that questions of taste or mood don't affect your choice. Let us also assume that each chef botches it, on occasion, so that you cannot solve the problem by straightforward exhaustive sampling.

[5] Although this assumption is not, strictly speaking, required.

[6] This was shown by Wei and Durham (1978).

in the sense that two agents interact and play a game if and only if they are connected by an edge.

For simplicity, let us assume that the underlying game is symmetric. (This ensures we do not need to worry whether a player takes the role of Row or Column, potentially having different strategy sets.) Each agent begins life with a single Pólya urn containing one ball of a unique color for each of her possible strategies.

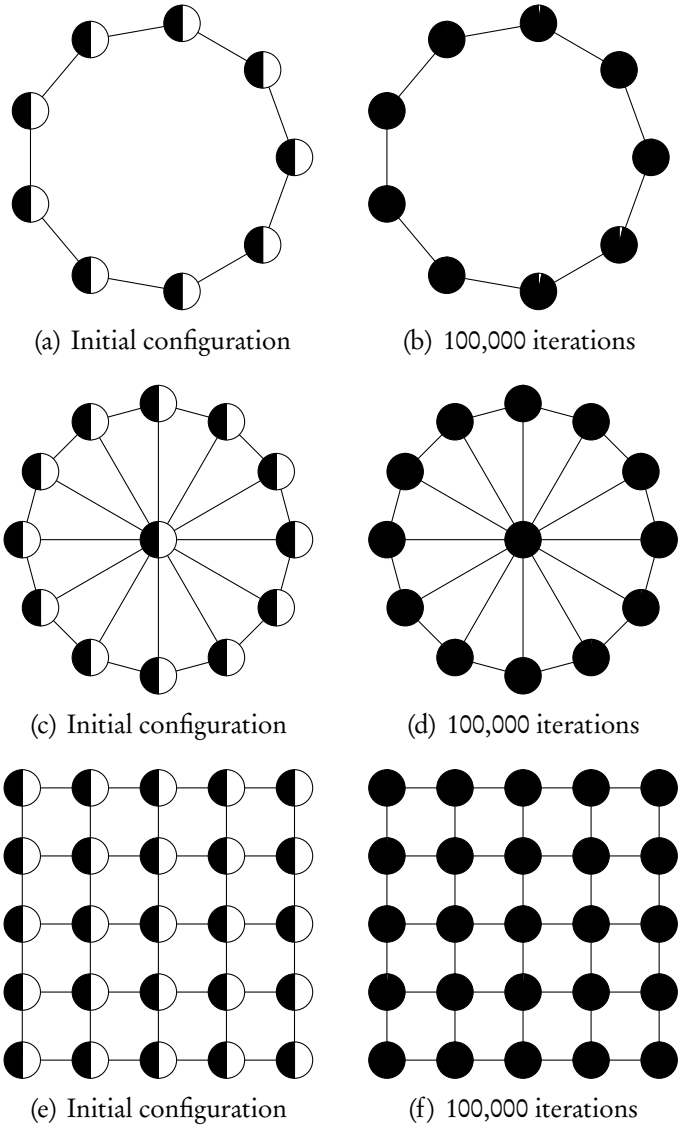
Each iteration, the pairwise interactions occur asynchronously and in a random order. When two agents interact, each reaches into his or her urn and draws a ball at random, with replacement. Each agent plays the strategy corresponding to the ball drawn from his or her urn, receiving a payoff. After the interaction, both agents reinforce by adding additional balls to their urn, the same colour as the one drawn, where the number of new balls added is determined by the payoff amount.[7]

Figure 1 illustrates the outcome of Roth-Erev reinforcement learners on three different social networks: the ring, wheel, and a grid. The underlying game was the canonical Prisoner's Dilemma with payoffs as indicated. The probability of agents choosing either Cooperate or Defect is displayed as a pie chart, with the white region representing the probability of cooperating and the black region representing the probability of defecting. Each action had an initial weight of 10, which prevented the outcome from the first round of play from severely skewing the probability of future actions.

This result is in accordance with the result of Beggs (2005), who showed that in a 2×2 game the probability a Roth-Erev reinforcement learner will play a strictly dominated strategy converges to zero. The one difference between this model and that of Beggs is that, here, the asynchronous dynamics allows two opponents to play a game with the collective urn configuration in a state not obtainable in Beggs's framework. That is, if a player A is connected to B and C by two edges, and A first interacts with B, then A — who will have reinforced after his interaction with B — may interact with C whose urn is in the same state as at the end of the previous iteration. However, as figure 1 illustrates, this has no real difference in the long-term convergence behaviour.

3. Cheap talk and reinforcement learning in networked games. In game theory, "cheap talk" refers to the possibility of players exchanging meaningless signals before choosing a strategy in a noncooperative game. Since players do not have the capability to make binding agreements, signal exchange was initially thought to be irrelevant for purposes of equilibrium selection in one-shot games. However, cheap-talk is more interesting than it might initially appear. In the case

[7] For simplicity, I assume that all payoffs are nonnegative integers.



(a) Initial configuration (b) 100,000 iterations

(c) Initial configuration (d) 100,000 iterations

(e) Initial configuration (f) 100,000 iterations

Figure 1: Effective convergence to Defect in the Prisoner's Dilemma played on three different structures. Payoff matrix: $T = 4, R = 3, P = 2, S = 1$ with an initial weight of 10 on the actions Cooperate and Defect. The nodes in the graph are pie charts showing a player's probability of choosing Cooperate (white) or Defect (black) from the urn.

of evolutionary game theory, (Skyrms, 2003, pp. 69–70) shows how cheap talk in the Stag Hunt creates a new evolutionarily stable state which does not exist in the absence of cheap talk.

Consider, then, an extension of the model presented in section 2 which incorporates a pre-play round of cheap talk on which players may condition their response. Since there seems little reason to restrict the number of signals a player may send, let's model the cheap-talk exchange using the method of signal invention from Skyrms (2010), based upon Hoppe-Pólya urns. Each player begins with a signalling urn containing a single black ball, known as the *mutator*.[8] When the mutator is drawn, the player chooses a new ball of a unique colour and sends that as the signal. Upon receipt of a signal, a player conditions her response upon the signal as follows: if this is the first time that the signal was received, the player creates a new response urn (a Pólya urn) labelled with that signal. The new response urn initially contains one ball of a unique colour for each strategy available to the player. A strategy is selected at random by sampling from the response urn with replacement. The game is played, after which reinforcement occurs; unlike the previous model, though, here both the signalling and response urns are reinforced, with the amount of reinforcement determined by the payoff. If the signal received by the player *had* been received previously, the player selects a strategy the same way, but uses the already-existing response urn labelled with that signal. Hence, the probability of choosing any particular strategy for a received signal will vary in a path-dependent way based on previous reinforcement.

Figure 2 illustrates aggregate results from 100 simulations for a simple cycle graph consisting of five agents. Cheap talk, here, makes essentially no difference in the long-term behaviour of the population: people still converge upon defection.[9] This should come as no surprise: the method of incorporating cheap talk means that a single individual, instead of playing the Prisoner's Dilemma with a single Pólya urn, can be thought of as being "partitioned" into several individuals each of whom play the Prisoner's Dilemma with their own Pólya urns. Since we know that Roth-Erev reinforcement learning (which is what the Pólya urn models) learns to avoid playing strictly dominated strategies in the limit, so will people using Roth-Erev reinforcement learning when they have the ability to conditionally respond to cheap talk.

4. Discounting, cheap talk, and the emergence of cooperation. It has been known for some time that models of cheap talk with signal invention often

[8]So called because Hoppe-Pólya urns were originally used as a model of neutral evolution.

[9]One might note that figure 2 still shows a frequency of cooperation of about 5% after 10,000 rounds of play. This apparent discrepancy is simply due to the smaller number of iterations involved.

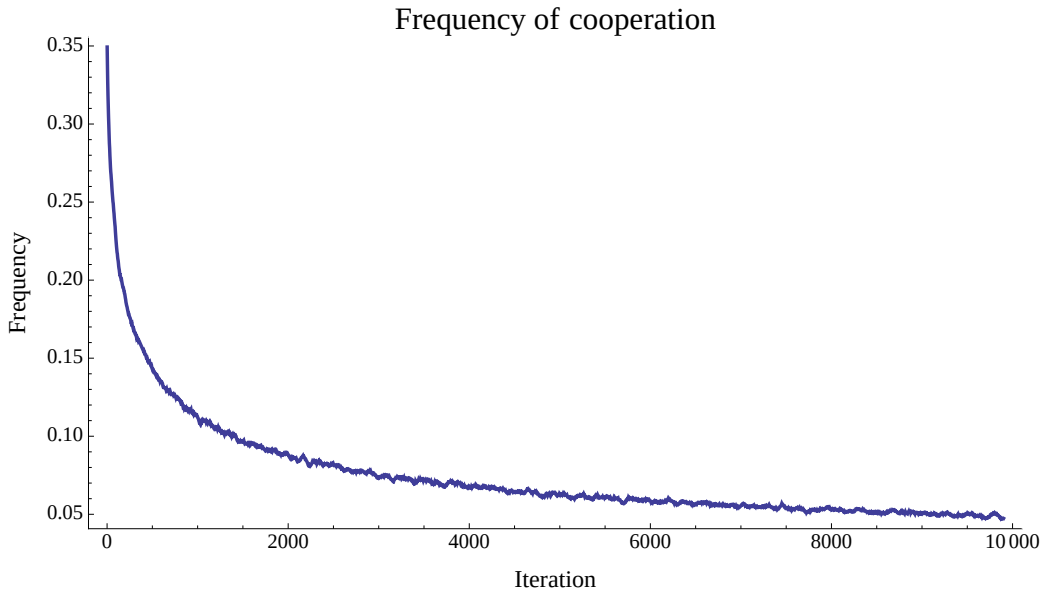
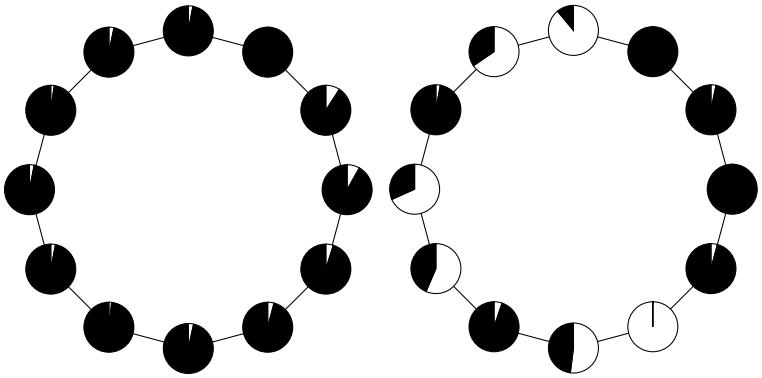


Figure 2: Aggregate results of 100 simulations featuring conditional response to cheap-talk and reinforcement learning, on a cyclic network with five agents. The payoff matrix for the Prisoner's Dilemma had $T = 4$, $R = 3$, $P = 2$ and $S = 1$.

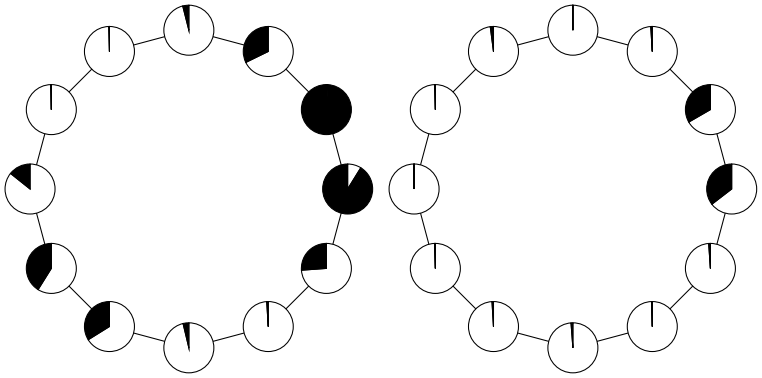
benefit from including a method of *pruning* the number of signals created. In Lewis sender-receiver games, for example, signal invention and reinforcement learning lead to efficient signalling systems, but with the side-effect of there being infinitely many signals in the limit (see Skyrms, 2010). However, Alexander et al. (2012) later showed that, if the model of signal invention and reinforcement learning is supplemented with signal "de-enforcement," efficient — and often *minimal* — signalling systems are produced. In a separate paper, also concerned with Lewis sender-receiver games, Alexander (forthcoming) showed that models of signal invention and reinforcement learning in which past information is discounted avoid excessive lock-in to particular signalling systems. This means that individuals are able to coordinate on efficient signalling systems yet, at the same time, respond rapidly to external stochastic shocks which change what the "correct" action is.

Consider, then, the model from the previous section with one final addition: each player has a discount factor δ that is applied to the weights of the signalling and response urns at the start of each iteration.[10] Since a seldomly-used signal will

[10]At this point, it becomes necessary to reinterpret the urn model. Instead of thinking of discrete balls in an urn, think instead of non-negative, real-valued numeric weights attached to signals (or strategies). The probability of selecting a signal (or strategy) to use is proportional to its weight after renormalisation; that is, let w_i denote the weight attached to signal (or strategy) i.



(a) 100 iterations (b) 1100 iterations



(c) 2100 iterations (d) 10,000 iterations

Figure 3: The emergence of cooperation in the Prisoner's Dilemma under cheap-talk, reinforcement learning, and discounting the past. The discount factor used was 0.95 and the cutoff threshold was 0.01.

have its weight eroded over time, let us introduce a *cutoff-threshold* τ such that, if the signal's weight drops below τ, it is eliminated entirely.[11] Finally, although the weights in the response urns are discounted, the cutoff threshold does not apply to them.[12]

Then the probability of selecting i is just $\frac{w_i}{\sum_j w_j}$. Discounting the past corresponds to multiplying each of the weights by the discount factor δ before reinforcement occurs.

[11]One technical complication lies with how to treat the mutator. If the mutator ball were eliminated, then signal invention would stop. Since there seems no principled reason to allow signal invention for only a short period of time (which is what would happen, since the mutator is never reinforced), in what follows it is assumed that the mutator is exempt from discounting.

[12]The reason why is as follows: strategies, here, stand for real, physical possibilities of action. One cannot simply eliminate a real, physical possibility in the same way one can eliminate an arbitrary constructed convention, like a signal.

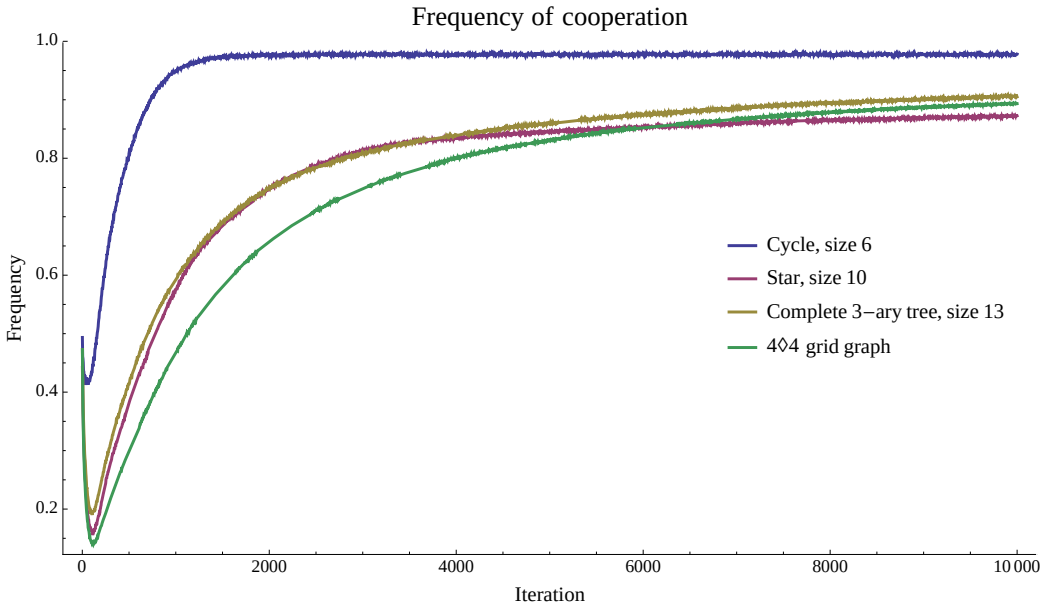


Figure 4: The emergence of cooperation, aggregate results for 1,000 simulations of 10,000 iterations each on a variety of networks.

With these adjustments to the model, we find a striking result: individuals rapidly move to defection, in the beginning, but then *learn to cooperate* over time. The combination of signal invention, reinforcement learning, and discounting the past enables the population to crawl out of the collectively suboptimal state in which they initially find themselves. Figure 3 illustrates this for a population of 12 agents on a cyclic network.

Figure 4 shows that the phenomenon of agents starting off with defection, and then learning to cooperate, occurs quite generally. Each line shows the aggregate results of 1,000 simulations, where the y-value at the x^{th} iteration is the frequency of cooperative acts across all 1,000 simulations at that iteration.[13] The graph topology has a notable effect on the speed with which cooperation emerges, but what is striking in light of the earlier results is how *often* cooperation happens. Recall that it was a *theorem* of Beggs (2005) that Roth-Erev reinforcement learning would play a strictly dominated strategy with a probability converging to zero in the limit.

[13]The values have been normalised to take into account the varying number of acts in a given iteration due to the graph topology. Since a cycle of size 6 has 12 actions each iteration (two per edge), the total number of cooperative acts at the x^{th} iteration was divided by $\frac{1}{12{,}000}$ to yield a value in the range $[0, 1]$. Likewise, the complete 3-ary tree of size 13 (with 24 actions per iteration), and the 4×4 grid graph (with 48 actions per iteration) had their aggregate values adjusted by factors of $\frac{1}{24{,}000}$ and $\frac{1}{48{,}000}$, respectively.

Why does cooperation emerge in the presence of discounting, but not otherwise? It seems to involve the following interaction of factors. Firstly, discounting places a cap on the overall weight a signal or action can receive as a result of reinforcement. For modest discount factors, say 95%, this means that there is a significant chance that a new signal will be attempted at any period in time. Secondly, suppose that a new signal is used between two agents, both of whom cooperate. If that signal (simply through chance) is used two or three times in a row, notice what happens: the amount of reinforcement in the default Prisoner's Dilemma adds 2 balls to both the signalling and response urn for the respective signal and action. If that happened, say, three times in a row, the weights attached to the other signals and responses would have been decreased by nearly 15%, on top of the fact that mutual Cooperation pays twice that of mutual Defection.[14] Thirdly, since the mutual Punishment payoff for the standard Prisoner's Dilemma used here awards 1 to each player, the *maximum* possible weight which could be attached to the previously used signal would be on the order of 20, since $\sum_{k=0}^{\infty}\left(\frac{19}{20}\right)^k = 20$, whereas the maximum possible weight for signals used to coordinate Cooperation would be on the order of 40.[15] Finally, when multiple signals are available to the sender, each signal will be used less often. Since discounting applies to each urn each iteration, that means the weights attached to actions in unused response urns are all discounted by the same amount each iteration. This doesn't affect the actual probability of any action in the urn being selected the next time the response urn is used, since $\frac{\delta^k \cdot w_i}{\sum_j \delta^k \cdot w_j} = \frac{w_i}{\sum_j w_j}$, but it does mean that the next amount of reinforcement will have more greater effect than it would otherwise have. The combination of these four factors, taken together, favour the emergence of cooperation.[16]

5. Conclusion. On the many roads leading to cooperation, Roth-Erev reinforcement learning seldom appeared in cases where the cooperative outcome required people to use a strictly dominated strategy. It has been shown here that if the basic mechanism of Roth-Erev reinforcement learning is supplemented by the psychologically plausible addition of temporal discounting of the past, cheap

[14] If $\delta = 0.95$, then $\delta^3 = 0.857375$.

[15] I say "on the order of" because the fact that the same signal, and response urn, may be used along multiple edges complicates matters. If a player is incident on m edges, then the maximum possible weight attached to a signal which is solely used for coordinating mutual Defection would be $20m$.

[16] Cheap talk with signal invention is an essential part of the story, for if the number of possible signals to send did not increase over time, then the fourth observation would not apply. This point is confirmed by simulations involving reinforcement learning, discounting the past, but no cheap talk with signal invention: there, players converge to Defect *very* quickly.

talk and signal invention, cooperation can regularly emerge in cases where that requires use of a strictly dominated strategy. Perhaps the most interesting and unexpected feature of this model is that, in the short term, individuals typically defect, but then, over time, eventually learn to cooperate. We have thus identified one formal mechanism which *suffices* to generate the following well-known social phenomenon: that people, albeit initially uncooperative, may, by means of repeated interactions over time, eventually engage in mutually beneficial cooperative behaviour even in the absence of formal institutions to establish, monitor, or police it.

References.

Alexander, J. McKenzie (2007). *The Structural Evolution of Morality*. Cambridge University Press.

— (forthcoming). "Learning to Signal in a Dynamic World". *The British Journal for the Philosophy of Science* .

Alexander, J. McKenzie, Brian Skyrms, and Sandy Zabell (2012). "Inventing New Signals". *Dynamic Games and Applications* 2, 1: 129–145.

Alexander, Jason and Brian Skyrms (1999). "Bargaining with Neighbors: Is Justice Contagious?" *Journal of Philosophy* 96, 11: 588–598.

Axelrod, Robert (1986). "An evolutionary approach to norms". *American Political Science Review* 80, 4: 1095–1111.

Beggs, A. (2005). "On the Convergence of Reinforcement Learning". *Journal of Economic Theory* 122: 1–36.

Bergstrom, Carl T. and Michael Lachmann (1997). "Signalling among relatives. I: Is costly signalling *too* costly?" *Phil. Trans. R. Soc. Lond. B* 352: 609–617.

— (1998). "Signaling among relatives. III: Talk is cheap". *Proc. Natl. Acad. Sci.* 95, 9: 5100–5105.

Bicchieri, Cristina (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bowles, Samuel and Herbert Gintis (2004). "The evolution of strong reciprocity: cooperation in heterogeneous populations". *Theoretical Population Biology* 65, 1: 17–28.

Boyd, Robert and Peter J. Richerson (1992). "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups". *Ethology and Sociobiology* 13: 171–195.

Bush, R. R. and F. Mosteller (1951). "A Mathematical Model for Simple Learning". *Psychological Review* 58: 313–323.

— (1955). *Stochastic Models for Learning*. New York: Wiley.

Ellison, G. (1993). "Learning, Local Interaction and Coordination". *Econometrica* 61: 1047–1071.

Gintis, Herbert (2000). "Classical Versus Evolutionary Game Theory". *Journal of Consciousness Studies* 7, 1–2: 300–304.

Huttegger, Simon M. and Kevin J. S. Zollman (2010). "Dynamic Stability and Basins of Attraction in the Sir Philip Sidney game". *Proceedings of the Royal Society of London B: Biological Sciences* 277, 1689: 1915–1922.

Lachmann, Michael and Carl T. Bergstrom (1998). "Signalling among Relatives II: Beyond the Tower of Babel". *Theoretical Population Biology* 54: 146–160.

Maynard Smith, John (1991). "Honest signalling: the Philip Sidney game". *Animal Behavior* 42: 1034–1035.

Nowak, Martin A. and Robert M. May (1993). "The Spatial Dilemmas of Evolution". *International Journal of Bifurcation and Chaos* 3, 1: 35–78.

Robson, Arthur J. (1990). "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake". *Journal of Theoretical Biology* 144: 379–396.

Roth, Alvin E. and Ido Erev (1995). "Learning in Extensive Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term". *Games and Economic Behavior* 8: 164–212.

Skyrms, Brian (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

— (2001). "The Stag Hunt". *Proceedings and Addresses of the APA* 75: 31–41.

— (2003). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.

— (2010). *Signals: Evolution, Learning, & Information*. Oxford University Press.

Sober, Elliot and David S. Wilson (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.

Trivers, Robert L. (1971). "The evolution of reciprocal altruism". *The Quarterly Review of Biology* 46: 35–57.

Wei, L. J. and S. Durham (1978). "The Randomized Play-the-Winner Rule in Medical Trials". *Journal of the American Statistical Association* 73, 364: 840–843.

Zahavi, A. (1975). "Mate Selection — Selection for a Handicap". *Journal of Theoretical Biology* 53: 205–214.

Zahavi, A. and A. Zahavi (1997). *The Handicap Principle*. Oxford University Press.